A Method of Successive Corrections of the Control Subspace in the Reduced-Order Variational Data Assimilation*

MAX YAREMCHUK

Department of Physics, University of New Orleans, New Orleans, Louisiana

DMITRI NECHAEV

Department of Marine Science, University of Southern Mississippi, Stennis Space Center, Mississippi

GLEB PANTELEEV

International Arctic Research Center, University of Alaska Fairbanks, Fairbanks, Alaska

(Manuscript received 13 March 2008, in final form 5 February 2009)

ABSTRACT

A version of the reduced control space four-dimensional variational method (R4DVAR) of data assimilation into numerical models is proposed. In contrast to the conventional 4DVAR schemes, the method does not require development of the tangent linear and adjoint codes for implementation. The proposed R4DVAR technique is based on minimization of the cost function in a sequence of low-dimensional subspaces of the control space. Performance of the method is demonstrated in a series of twin-data assimilation experiments into a nonlinear quasigeostrophic model utilized as a strong constraint. When the adjoint code is stable, R4DVAR's convergence rate is comparable to that of the standard 4DVAR algorithm. In the presence of strong instabilities in the direct model, R4DVAR works better than 4DVAR whose performance is deteriorated because of the breakdown of the tangent linear approximation. Comparison of the 4DVAR and R4DVAR also shows that R4DVAR becomes advantageous when observations are sparse and noisy.

1. Introduction

In the past two decades, the methods of oceanographic data assimilation into numerical models have undergone a significant progress from the early works of Le Dimet and Talagrand (1986) and Thacker (1988) to solution of the increasingly complex problems, reflected in a series of monographs by Bennett (1992), Wunsch (1996), Evensen (2006), and Talagrand and Bouttier (2009), among others.

Most recently, research in data assimilation has an apparent trend toward the studies of the ensemble-based sequential techniques (Evensen 2003; Ott et al. 2004; Zupanski 2005; Uzunoglu et al. 2007). These methods utilize low-dimensional ensembles of model states to approximate propagation of error covariances that are vital for improvement of practical weather forecast. At the same time, the classic strong constraint four-dimensional variational data assimilation (4DVAR) methods still remain an important tool in atmospheric and oceanic data analysis in both global (Wenzel et al. 2001; Stammer et al. 2003; Blessing et al. 2008) and regional (Zupanski et al. 2005; Yaremchuk 2006; Di Lorenzo et al. 2007) applications. The strong constraint methods are of particular importance in oceanography where the data coverage is sparse and observations are less accurate.

With the ever-growing complexity and resolution of the ocean general circulation models (OGCMs), constraining them by 4DVAR methods is hampered by the following difficulties:

1) *High computational cost of 4DVAR optimization.* OGCMs have the typical state vector dimension of

* International Pacific Research Center Contribution Number 583 and School of Ocean and Earth Science and Technology Contribution Number 7630.

Corresponding author address: Max Yaremchuk, Navy Research Laboratory, Code 7321, Bldg. 1009, Stennis Space Center, MS 39529.
E-mail: myaremch@uno.edu

DOI: 10.1175/2009MWR2592.1

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
FEB 2009		**00-00-2009 to 00-00-2009**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
A Method of Successive Corrections of the Control Subspace in the Reduced-Order Variational Data Assimilation	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Naval Research Laboratory,Code 7321,Stennis Space Center,MS,39529**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

A version of the reduced control space four-dimensional variational method (R4DVAR) of data assimilation into numerical models is proposed. In contrast to the conventional 4DVAR schemes, the method does not require development of the tangent linear and adjoint codes for implementation. The proposed R4DVAR technique is based on minimization of the cost function in a sequence of low-dimensional subspaces of the control space. Performance of the method is demonstrated in a series of twin-data assimilation experiments into a nonlinear quasigeostrophic model utilized as a strong constraint. When the adjoint code is stable, R4DVAR?s convergence rate is comparable to that of the standard 4DVAR algorithm. In the presence of strong instabilities in the direct model, R4DVAR works better than 4DVAR whose performance is deteriorated because of the breakdown of the tangent linear approximation. Comparison of the 4DVAR and R4DVAR also shows that R4DVAR becomes advantageous when observations are sparse and noisy.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**13**	

10^6–10^7 whereas the number of independent observations is only an order in magnitude smaller and growing. On the other hand, optimal estimation of an ocean state with classical 4DVAR methods requires the number of model runs [including those of the tangent linear (TL) and/or adjoint models] comparable with the number of observations or model state dimension, that is computationally prohibitive. As a consequence, applications of 4DVAR methods are limited to finding suboptimal solutions, obtained after 10–100 iterations of the minimization procedure.

2) *High maintenance cost of the adjoint and tangent linear codes.* A significant part of the programming burden related to the maintenance of the adjoint and TL codes cannot be automated at the present state of the adjoint compilers. In addition, keeping the adjoint and TL codes updated in parallel with the perpetually upgraded models is labor intensive and prone to human errors.
3) *Breakdown of the tangent linear approximation (TLA).* In the presence of strong physical instabilities of the background state applicability of TLA is restricted to relatively short time intervals (e.g., Oldenborgh et al. 1999). Furthermore, the TL and adjoint codes of the community OGCMs never represent exact TL or adjoint operators, especially when model physics contains parameterized discontinuities (Zhu et al. 2002).

To resolve these difficulties the focus of research has recently shifted toward the development of the reduced control space 4DVAR (R4DVAR) methods. As a few examples, Robert et al. (2005) utilized the empirical orthogonal (EOF) analysis of the model trajectory for the definition of the reduced control space and parameterization of the background error covariance. Robert et al. (2006) explored preconditioning of the incremental 4DVAR assimilation by the R4DVAR method. Qiu et al. (2007) studied a possibility to use an ensemble of randomly perturbed model states for generation of the reduced control by singular value decomposition. Another strategy studied by (Cao et al. 2007; Daescu and Navon 2007) is based on the reduction of the model itself using EOF approach. Although the latter technique improves computational efficiency, the issue of finding an optimal low-dimensional state subspace remains an open question.

This paper presents a version of the reduced control space 4DVAR data assimilation method. In contrast to previous studies (e.g., Robert et al. 2006; Daescu and Navon 2007; Qiu et al. 2007; Liu et al. 2008), which utilize a *fixed* EOF-generated subspace for optimization, our algorithm employs a sequence of low-dimensional subspaces that are iteratively updated in the process of finding a minimum of the cost function.

The paper is organized as follows: in the next section we first present linear considerations, underlying the development of the scheme and outline the algorithm. In section 3 the setup of the twin-data experiments is described and the issue of the adjoint code instability is considered. In section 4 we present the results of the twin-data experiments and compare the method with the standard 4DVAR technique. The conclusions are presented in section 5.

2. Linear background

In a linear context, a 4DVAR iterative procedure employs the adjoint code for exact multiplication of an arbitrary vector by the Hessian matrix. For computational reasons, however, the number of iterations in practical problems is often limited by a few hundred. Therefore, such solutions should be treated as optimizations on the subspace spanned by a limited number of eigenvectors of the Hessian matrix.

In this regard, the ability to retrieve leading eigenvectors of the Hessian matrix is of primary importance for practical purposes. In this section, we consider a simplified linear variational data assimilation problem controlled by the initial conditions, employ EOF analysis of the model solutions to obtain low-dimensional approximations to the Hessian structure, and construct a minimization method, which does not require the adjoint algorithm.

a. EOF analysis of model solutions

As discussed by Farrell and Ioannou (2001, 2006), an optimal reduction of a dissipative normal dynamical system can be represented as Galerkin projection of the dynamics onto the least damped modes. In this study, we consider control of model solutions by the initial conditions. Therefore it will be instructive to consider first the impact of initial conditions on the result of EOF analysis of the corresponding model solution in the simplified case of the normal dynamical operator.

Consider a model of oceanic circulation $\partial_t \mathbf{x} = \mathbf{M}\mathbf{x}$, where $\mathbf{x}$ is the vector of the state variables and $\mathbf{M}$ is a normal differential operator with a full set of eigenfunctions μ_k and corresponding eigenvalues λ_k, which satisfy the conditions $\mathrm{Re}\{\lambda_k\} \leq 0$ and $\tilde{\lambda} = \min_k |\mathrm{Im}\lambda_k| > 0$. If $\mathbf{M}$ is time independent, a model solution corresponding to the initial state $\mathbf{x}^0$ is $\mathbf{x}(t) = \mathbf{x}^0 \exp(\mathbf{M}t)$. In terms of μ_k the solution is represented by the expansion $\mathbf{x}(t) = \Sigma_k a_k \exp(\lambda_k t)\mu_k$, where a_k are the projections of $\mathbf{x}^0$ on the eigenstates of $\mathbf{M}$.

A standard technique widely used for initialization of the sequential data assimilation schemes (e.g., Robert et al. 2005) is the EOF analysis of the covariance matrix $\mathbf{C}$ generated by the time averaging of a model run over a sufficiently long-time interval T. When written in terms of $\mu_k(x)$, the covariance matrix is

$$\mathbf{C} \equiv \overline{\mathbf{x}(t)\mathbf{x}^{\mathrm{T}}(t)} = \frac{1}{T}\int_0^T \sum_{k,l} a_k a_l^* \exp[(\lambda_k + \lambda_l^*)t]\mu_k \mu_l^{*\mathrm{T}}\, dt = \sum_{k,l} a_k a_l \frac{\exp[(\lambda_k + \lambda_l^*)T] - 1}{(\lambda_k + \lambda_l^*)T}\mu_k \mu_l^{*\mathrm{T}},$$

where the $^{\mathrm{T}}$ symbol denotes transposition and the asterisk stands for the complex conjugate. If the averaging time is long enough ($\tilde{\lambda}T \to \infty$), the off-diagonal elements of $\mathbf{C}$ vanish in the basis $\{\mu_k\}$ and its largest eigenvalue Λ_m, satisfying the condition $\mathrm{Re}\{\lambda_m\}/\tilde{\lambda} \to 0$, can be estimated as

$$\Lambda_m = |a_m|^2 \frac{\exp(2\mathrm{Re}\{\lambda_m\}T) - 1}{2\mathrm{Re}\{\lambda_m\}T} \equiv |a_m|^2 F(\mathrm{Re}\{\lambda_m\}T). \tag{1}$$

The expression in (1) shows that the leading eigenvalues of $\mathbf{C}$ are the squares of the largest μ components of $\mathbf{x}^0$ corresponding to eigenstates of $\mathbf{M}$ with the smallest damping $\mathrm{Re}\{\lambda\}$. Since dissipation in $\mathbf{M}$ usually selectively damps high-frequency modes, the leading eigenvectors of $\mathbf{C}$ tend to capture the largest spatial scales of model variability, which is also typical for the results of EOF analyses of the oceanic data. At the same time, the largest spatial scales are the least prone to parameterization errors of the dissipative processes in OGCMs. This property makes the EOF decomposition of a model solution an efficient tool for selectively retrieving those components of $\mathbf{x}^0$ that are most accurately projected by a numerical model on the data points distributed over a given time interval.

b. *Krylov subspace methods and the large least squares problems*

Now consider a problem of 4DVAR into a linear dynamical system $\partial_t \mathbf{x} = \mathbf{M}\mathbf{x}$, where $\mathbf{x} \in \mathcal{R}^M$ is the state vector of the ocean and $\mathbf{M}$ is a time-dependent linear operator. The model solutions are controlled by the initial state $\mathbf{x}^0 \equiv \mathbf{x}(t^0)$. Observations d^n of $\mathbf{x}$ are made at times t^n, $n = 0, \ldots, N$, with $N \ll M$. We adopt a linear model for observations $d^n = \mathcal{O}_n\mathbf{x}(t^n) + \varepsilon$, where ε is the spatially uncorrelated noise with the inverse covariance R_n and linear operators $\mathcal{O}_n$ project the model states $\mathbf{x}(t^n)$ onto the observed quantities d^n. Since the total number of observations in the time interval $[t^0, t^N]$ is usually much less than M, a background model state $\mathbf{x}_b^0$ and its inverse error covariance $\mathbf{B}$ are utilized for regularization of the problem.

Introducing notations $\mathcal{G}_n = \sqrt{\mathcal{R}}_n\mathcal{O}_n$, $\overline{d}^n = \sqrt{R}_n d^n$ and $\mathbf{A}^n$ for the propagator between t^0 and t^n, the standard formulation of the 4DVAR assimilation problem (e.g., Bennett 1992) can be written down as

$$J = \frac{1}{2}\left[(\mathbf{x}^0 - \mathbf{x}_b^0)^{\mathrm{T}}\mathbf{B}(\mathbf{x}^0 - \mathbf{x}_b^0) + \sum_n (\mathcal{G}_n\mathbf{A}^n\mathbf{x}^0 - \overline{d}^n)^{\mathrm{T}}(\mathcal{G}_n\mathbf{A}^n\mathbf{x}^0 - \overline{d}^n)\right] \to \min_{\mathbf{x}^0}. \tag{2}$$

Minimization of J is can be reduced to solution of the normal equation:

$$\frac{\partial J}{\partial \mathbf{x}^0} = \left(\mathbf{B} + \sum_n \mathbf{A}^{n\mathrm{T}}\mathcal{G}_n^{\mathrm{T}}\mathcal{G}_n\mathbf{A}^n\right)\mathbf{x}^0 - \left(\mathbf{B}\mathbf{x}_b^0 + \sum_n \mathbf{A}^{n\mathrm{T}}\mathcal{G}_n^{\mathrm{T}}\overline{d}^n\right) = 0, \tag{3}$$

which can be rewritten as $\mathbf{H}\mathbf{x}^0 = \mathbf{b}$, where $\mathbf{b} = \mathbf{B}\mathbf{x}_b^0 + \sum_n \mathbf{A}^{n\mathrm{T}}\mathcal{O}_n^{\mathrm{T}}\mathcal{R}_n\overline{d}^n$ and

$$\mathbf{H} = \frac{\partial^2 J}{\partial(\mathbf{x}^0)^2} = \mathbf{B} + \sum_n \mathbf{A}^{n\mathrm{T}}\mathcal{G}_n^{\mathrm{T}}\mathcal{G}_n\mathbf{A}^n \tag{4}$$

is the Hessian matrix. Assuming that the symmetric $M \times M$ matrix $\mathbf{B}$ could be represented as $\mathbf{B} = \mathbf{Q}^{\mathrm{T}}\mathbf{Q}$, it is convenient to rewrite the minimization problem (2) in a symmetric form:

$$J = \frac{1}{2}(\mathbf{S}\mathbf{x}^0 - \boldsymbol{d})^{\mathrm{T}}(\mathbf{S}\mathbf{x}^0 - \mathbf{d}), \tag{5}$$

where

$$\mathbf{S} = \begin{bmatrix} \mathbf{Q} \\ \mathcal{G}_1\mathbf{A} \\ \vdots \\ \mathcal{G}_N\mathbf{A}^N \end{bmatrix}, \quad \mathbf{d} = \begin{bmatrix} \mathbf{Q}\mathbf{x}_b^0 \\ \overline{d}_1 \\ \vdots \\ \overline{d}^N \end{bmatrix} \tag{6}$$

are the "square root" of $\mathbf{H} \equiv \mathbf{S}^{\mathrm{T}}\mathbf{S}$ and the normalized data vector, respectively.

The large and sparse system of linear equations (3) could be solved by means of the Krylov subspace methods that form an orthogonal basis $\{\mathbf{e}_m\}$ on the sequence $\mathbf{H}^m\mathbf{r}_0$, $m = 1, \ldots, M$, where $\mathbf{r}_0 = \mathbf{H}\mathbf{x}_0^0 - \mathbf{b}$ is an arbitrary initial residual vector. The approximations to the solution are obtained by minimizing the residual over the Krylov subspaces $\mathcal{K}^m$ spanned by $\{\mathbf{e}_m\}$. The well-known generalized minimum residuals (GMRES),

conjugate gradient, and biconjugate gradient methods can be viewed as particular applications of the Krylov subspace technique (e.g., Saad 2003).

As is seen from (4), multiplication of the arbitrary residual vector by the Hessian matrix requires an algorithm for multiplication of the vector by $\mathbf{A}^{n\mathrm{T}}$ (the adjoint model). The latter may often be unavailable and/or expensive to run and implement (see section 1). Therefore, it might be useful to explore a possibility of constructing an ''adjointless'' iterative scheme for the minimization problem in (2), which is based on the Krylov subspace technique.

In this regard, inspection of (4) shows that computation of the higher powers of $\mathbf{H}$ for construction of the Krylov subspaces may seem to be redundant, because, for example, $\mathbf{H}^M$ contains terms with powers of $\mathbf{A}$ up to $2NM$ whereas a complete basis in $\mathcal{R}^M$ could, in principle, be built on the sequences $\{\mathbf{A}^{k\mathrm{T}}\mathcal{G}_k^{\mathrm{T}}\mathcal{G}_k\mathbf{A}^k\mathbf{r}\}$, $\{\mathcal{G}_k\mathbf{A}^k\mathbf{r}\}$, or $\{\mathbf{A}^k\mathbf{r}\}$ with $k = 1, \ldots, M$ (assume, for a moment, that both $\mathbf{A}$ and $\mathcal{G}_k^{\mathrm{T}}\mathcal{G}_k$ have the full rank). This observation gives some grounds to examine numerical data assimilation schemes, which do not require multiplication by $\mathbf{A}^{\mathrm{T}}$ in the construction of Krylov subspaces.

The simplest approach is to build $\mathcal{K}^m$ on the powers of $\mathbf{A}$. Although this method does not explicitly take into account the structure of $\mathbf{B}$ and $\mathcal{G}_k$, it tends to selectively extract the least damped components of $\mathbf{r}$ (i.e., those components that are most accurately projected by $\mathbf{A}^k$ on the data; section 2a). This property could be achieved by extracting the leading eigenvectors of the covariance matrix $\mathbf{C} = \Sigma_n(\mathbf{A}^n\mathbf{r})(\mathbf{A}^n\mathbf{r})^{\mathrm{T}}$ through the EOF analysis of its dual $\mathbf{C}^*(i, k) = \langle \mathbf{A}^i\mathbf{r}, \mathbf{A}^k\mathbf{r}\rangle$, where angular brackets denote inner product $\mathcal{R}^M$.

The second adjointless approach is to build $\mathcal{K}^m$ on the sequence $\{\mathcal{G}_n\mathbf{A}^n\mathbf{r}\}$. This method requires a definition of the operators $\tilde{\mathcal{P}}_n$ to project the model counterparts of the data on the state space at times t^n. Inspection of (4) shows that specifying $\tilde{\mathcal{P}}_n = \mathbf{A}^{n\mathrm{T}}\mathcal{G}_n^{\mathrm{T}}$ provides a ''natural'' scheme $\mathcal{K}^k = \mathrm{span}\{\mathbf{A}^{k\mathrm{T}}\ \mathcal{G}_k\mathcal{G}_k\mathbf{A}^k\mathbf{r}\}$. This method, however, should be discarded, as it contains $\mathbf{A}^{\mathrm{T}}$. Mathematically, there is a considerable freedom in defining the projection operators: the only restriction imposed on $\tilde{\mathcal{P}}_n$ by the requirement of convergence of the Krylov scheme to the solution of (2) is to keep the rank of the modified $\mathbf{S}^{\mathrm{T}}$ intact (Hayami et al. 2007). This can be achieved, for example, by replacing $\mathbf{A}^{\mathrm{T}}$ by $\mathbf{A}$ in the expression for $\mathbf{S}^{\mathrm{T}}$. Computationally, such a replacement will require an additional model run (as a substitution of the adjoint) for the generation of $\mathcal{K}^n$. In the present study, we take another approach and utilize the background error covariance $\mathbf{B}^{-1}$ for projection by setting $\tilde{\mathcal{P}}_n = (\mathcal{D}_n^{\mathrm{T}}\mathcal{D}_n)^{-1}\mathcal{G}_n^{\mathrm{T}}$ with $\mathcal{D}_n^{\mathrm{T}} = [\mathcal{G}_n^{\mathrm{T}}, \mathbf{Q}^{\mathrm{T}}]$. This choice could be supported by the fact that the covariance propagates in time by model dynamics and thus provides a measure for the distance between the model states $\mathbf{x}(t)$ in terms of their value at $t = 0$.

One can expect that this second method of generating $\mathcal{K}^k$ may converge faster, because it takes into account the structure of $\mathcal{G}_k$ and $\mathbf{B}$ in generating the Krylov spaces and, therefore, may give better approximations to $\mathbf{H}$ than the first method.

c. Practical implementation

In this section we describe a 4DVAR assimilation method based on successive minimizations of the cost function performed in low-dimensional Krylov subspaces $\mathcal{K}^m$ spanned by projections of the residuals on the approximations to the leading eigenmodes of $\mathbf{H}$ and $\mathbf{C}$ in different experiments. The proposed technique exploits low computational cost of both EOF analysis and explicit inversions of the Hessian operators in $\mathcal{K}^m$.

The optimization procedure starts with a first-guess state $\mathbf{x}_0^0$, whose time evolution is subsampled to retrieve the first Krylov space $\mathcal{K}_0^m$ via EOF decomposition of the corresponding covariance matrix $\mathbf{C}$. The corresponding projection operator $\mathcal{P}_0$ is represented by the $M \times m$ matrix, whose columns are the eigenvectors $\{\mu_i^0\}$ of $\mathbf{C}$. The dimension m of $\mathcal{K}^m$ is somewhat arbitrary, but should be small enough to reduce the computational cost. Objectively, m can be chosen, for example, as the number of modes, explaining a certain portion $1 - \xi$ of the model variability defined by the observational noise level ξ.

After specifying the first Krylov subspace, the gradient $\nabla_0 J$ in $\mathcal{K}_0^m$ is computed by perturbing $\mathbf{x}_0^0$ with $\{\mu_i^0\}$ and taking the finite differences of J. Simultaneously, we obtain the projection of the Hessian operator $\mathbf{H}_0 = \mathcal{P}_0^{\mathrm{T}}\mathbf{H}\mathcal{P}_0$ in $\mathcal{K}_0$ using the approach of Zupanski (2005). Next, the value of control $\mathbf{x}_1^0$ after the first iteration is computed as

$$\mathbf{x}_1^0 = \mathbf{x}_0^0 - \tilde{\mathbf{x}}_0^0 \equiv \mathbf{x}_0^0 - \mathcal{P}_0\mathbf{H}_0^{-1}\nabla_0 J. \tag{7}$$

Note that in the case of linear dynamics, $\mathbf{x}_1^0$ corresponds to the exact and unique minimum of J in $\mathcal{K}_0^m$. In the nonlinear case considered in the next section, it is necessary to execute several iterations of this ''internal'' optimization loop to reach a local minimum.

The second ''external'' iteration starts with the EOF analysis of $\mathbf{x}_1(t)$, which generates the next Krylov subspace $\mathcal{K}_1^m$. Note that the residual control $\mathbf{x}_1^0$ does not contain the $\mathcal{K}_0$ components of $\mathbf{x}_0^0$ (denoted by $\tilde{\mathbf{x}}_0^0$), which already explain a certain portion of the data. To remove them from $\mathcal{K}_1^m$, the basis in $\mathcal{K}_1^m \times \mathcal{K}_0^m$ is orthogonalized using the Gram–Schmidt process.

Next, we store the suboptimal control $\tilde{\mathbf{x}}^0 = \tilde{\mathbf{x}}_0^0$ and exclude it from optimization process by updating the cost function in (5) and minimizing it in $\mathcal{K}_1^m$:

$$J \to J = \frac{1}{2}[\mathbf{S}(\mathbf{x}^0 + \tilde{\mathbf{x}}^0) - \mathbf{d}]^{\mathrm{T}}[\mathbf{S}(\mathbf{x}^0 + \tilde{\mathbf{x}}^0) - \mathbf{d}] \to \min_{\mathbf{x}^0 \in \mathcal{K}_1^m}. \tag{8}$$

Minimization in $\mathcal{K}_1^m$ is performed by the procedure outlined by (7). In a similar manner we proceed with further iterations.

In general, on the ith iteration, we first update the suboptimal control $\tilde{\mathbf{x}}^0 \leftarrow \tilde{\mathbf{x}}^0 + \tilde{\mathbf{x}}_i^0$, find $\mathcal{K}_{i+1}$ through the EOF analysis of $\mathrm{x}_i(t)$, orthogonalize the basis in $\mathcal{K}_{i+1}^m \times \mathcal{K}_i^m$, update the cost function in (8), and then minimize it in $\mathcal{K}_{i+1}^m$ to obtain the next contribution $\tilde{\mathbf{x}}_{i+1}^0$ to the suboptimal control vector $\tilde{\mathbf{x}}^0$. Numerically, $\tilde{\mathbf{x}}_{\mathrm{i}}^0$ are iteratively accumulated in a single array $\tilde{\mathbf{x}}^0$ and do not require additional memory resources.

d. Comparison with GMRES

The proposed algorithm belongs to the family of Krylov subspace methods widely used for solution of the large sparse linear systems of equations. Therefore it is instructive to compare it with the well-known limited-memory generalized minimal residual (GMRES$_m$) scheme (Saad 2003). When applied to the system of equations $\mathbf{H}\mathbf{x} = \mathbf{b}$, the GMRES$_m$ algorithm [$m = \sup(\dim \mathcal{K})$] employs the following iterative loop:

1) Given the approximate solution $\tilde{\mathbf{x}}$ on the kth iteration, generate Krylov spaces spanned by $\{\mathcal{P}^i[\mathbf{H}]\mathbf{r}\}$, $i = 1, \ldots, k$, where $P^i[\mathbf{H}]$ are the ith-order polynomials in $\mathbf{H}$ and

$$\mathbf{r} = \mathbf{b} - \mathbf{H}\tilde{\mathbf{x}} \tag{9}$$

   is the residual. The polynomials $\mathcal{P}^i$ are generated sequentially by the Arnoldi process (Arnoldi 1951) to form orthonormal bases in $\mathcal{K}^i$.
2) Compute

$$\xi: \quad J = (\mathbf{H}\xi - \mathbf{r})^{\mathrm{T}}(\mathbf{H}\xi - \mathbf{r}) \to \min_{\xi \in \mathcal{K}^i} \tag{10}$$

   and update the approximation to the solution: $\tilde{\mathbf{x}} \leftarrow \tilde{\mathbf{x}} + \xi$.
3) If the updated residual is small enough, exit; otherwise, go to step 1 and either increase the dimension of the Krylov subspace by generating the next-order polynomial, or (if $i = m$) start building the new sequence of $\mathcal{K}^i$ using the updated residual.

It can be proved that if $m = M$, the GMRES algorithm provides the exact solution (Saad 2003). Moreover, the Krylov space can be built on the powers of any matrix, whose null space is $\mathbf{H}$-orthogonal to $\mathbf{b}$.

The proposed R4DVAR algorithm has two major differences from the classic GMRES$_m$. Both of them are dictated by the necessity to avoid multiplication by $\mathbf{A}^{\mathrm{T}}$. First, the Krylov spaces are built not on the powers of $\mathbf{H}$, but on the sequences of the operators approximating the entries of its square root in (6). Second, since computation of the residual in (9) requires the adjoint code, we adopt an alternative expression $\mathbf{r} = \mathbf{x}_0 - \tilde{\mathbf{x}}$ implicitly multiplying (9) by $\mathbf{H}^{-1}$. This modification does not affect the convergence properties of the algorithm as soon as the first-guess vector $\mathbf{x}_0$ contains all the spectral components of $\mathbf{H}$ that are needed for decomposition of $\mathbf{b}$.

Other distinctions from the classic GMRES$_m$ scheme are purely technical:

1) Orthogonalization of the basis in $\mathcal{K}^m$ is done not by the Arnoldi process, but via EOF analysis of the sample covariance matrix built on the Krylov vectors $\{\mathbf{A}^i\mathbf{r}\}$ or $\{\tilde{\mathcal{P}}^i\mathcal{G}_i\mathbf{A}^i\mathbf{r}\}$.
2) Minimization in the Krylov subspace is done by direct computation of the gradients and inversion of the Hessian in $\mathcal{K}^m$. From the computational point of view this is approximately equivalent to the GMRES minimization scheme, which employs the Gramm matrix generated by the Arnoldi process.
3) The residual cost function (10) is taken in its original form (8), which can be considered as a square root form of (10). This allows us to estimate the cost function by summing the squares of the residuals in the data space and thus avoid utilization of the adjoint code, which is necessary for estimation of (10).

From the mathematical point of view the proposed algorithm should be equivalent to full ($m = M$) GMRES under two conditions: 1) the rank of $\mathbf{S}^{\mathrm{T}}$ is kept intact by $\tilde{P}_k$; and 2) the first-guess vector contains all the spectral components of $\mathbf{b}$. These conditions do not seem to be too restrictive in applications, because a relatively good first-guess approximation is often available in the form of the background state $\mathbf{x}_b^0$. Note, however, that in contrast to full GMRES, which may work with *any* first-guess vector, the proposed algorithm relies on the quality of $\mathbf{x}^0$.

In the numerical experiments below we demonstrate the algorithm's performance in a typical "oceanographic application" when neither the background state nor its error covariance is available. In such situations the "best" first-guess state has to be retrieved from the data, the background state is taken to be zero, and its error covariance is modeled by a low-pass filter. This approach to error covariance modeling has gained considerable attention in recent years (e.g., Weaver and Courtier 2001; Pannekoucke and Massart 2008).

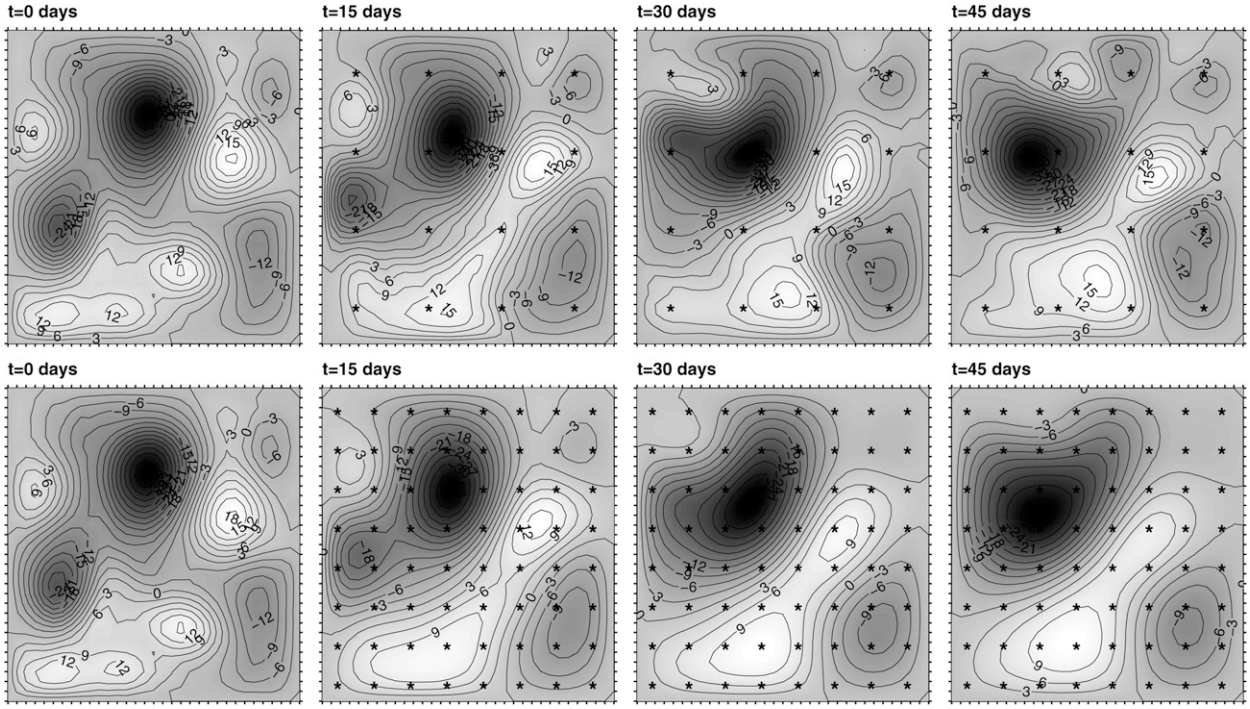

FIG. 1. Streamfunction ψ of the reference solutions with (top) $\nu = 50$ m^2 s^{-1} and (bottom) $\nu = 500$ m^2 s^{-1}. Asterisks denote the data points of the sampling grids used in twin-data experiments. The contour interval is 3000 m^2 s^{-1}.

3. Twin-data experiments

To assess the performance of the method, we conducted twin-data experiments with a nonlinear model controlled by the initial conditions. The underlying idea is to generate reference model solutions, sample them using a simulated ''observational array,'' contaminate the samples by noise, and then reconstruct the reference solution from these samples using the assimilation method under study (hereafter R4DVAR).

A nonlinear model is chosen for two reasons: 1) it is more realistic than the linear one in the sense of applications, and 2) it has an ability to generate reference solutions with unstable tangent linear and adjoint models. The latter situation is quite common in practice, and it was interesting to compare R4DVAR with the standard 4DVAR in that case.

a. Numerical model

We consider a quasigeostrophic model in a square 33×33 grid Ω with a spatial resolution of $\delta x = 15$ km (Fig. 1):

$$\partial_t q + J(\psi, \Delta\psi) + \beta\partial_x\psi = \nu\Delta^2\psi + \frac{1}{h}\mathrm{curl}_z\tau, \quad (11)$$

$$\Delta\psi - R_d^{-2}\psi = q, \quad (12)$$

where ψ is the streamfunction in the upper layer, β is the meridional gradient of the Coriolis parameter, R_d is the internal Rossby radius of deformation, and ν is the horizontal diffusion coefficient.

At the spinup stage the model is forced for 1000 days by a steady wind stress curl pattern:

$$\mathrm{curl}_z\tau = \frac{\tau_0}{L}\sin\left(4\pi\frac{x^*}{L}\right)\cos\left(4\frac{y^*}{L}\right).$$

Here $\tau_0 = 5 \times 10^{-5}$ m^2 s^{-2}, $L = 480$ km is the horizontal size of the domain, and x^*, y^* are Cartesian coordinates rotated 40° with respect to the north–south direction. The other model parameters are $h = 700$ m, $\beta = 2 \times 10^{-1}$ m^{-1} s^{-1}, and $R_d = 25$ km. The horizontal diffusion coefficient ν was either 50 or 500 m^2 s^{-1} in different experiments.

Since the boundary conditions are assumed to be known ($\psi|_{\partial\Omega} = \Delta\psi|_{\partial\Omega} = 0$), the model is controlled by the initial distribution of the potential vorticity field $q(x, y, 0)$. Equations (11) and (12) are integrated in time using the leapfrog scheme with a time step of 0.05 days. Therefore, the number of adjusted parameters M (gridpoint values of q at $t = 1000$ and $t = 1000.05$ days) is $2 \times 31^2 = 1922$.

After the spinup the wind was switched off and the model was run in an unforced regime for $T = 45$ days,

producing the reference solutions ψ^{ref} (Fig. 1) for the twin-data experiments.

b. Instability of the tangent linear model

The reference solution with $\nu = 50\ \text{m}^2\ \text{s}^{-1}$ is characterized by a high degree of nonlinearity (the typical value of $|J(\psi, \omega)|$ exceeds dissipation and β-effect terms by an order in magnitude). As a consequence, TL and adjoint codes turn to be unstable. This instability has an e-folding time scale τ_e of approximately 15 days, that is several times less than the typical dissipation time $\tau_d = \delta x^2/\nu$ of the grid-scale harmonics. This property of the model significantly degrades the performance of the traditional 4DVAR scheme, based on the adjoint code. We tested the validity of TLA by perturbing q at $t = 0$ with a test function:

$$\delta_q(x, y) = \varepsilon \sin\left(13\pi \frac{x}{L}\right) \sin\left(11\pi \frac{y}{L}\right)$$

and estimated the quantity:

$$\Phi(\varepsilon) = \frac{\left\|\psi_{q+\delta q} - \psi_q - \overline{\psi}^q_{\delta q}\right\|}{\left\|\psi_q\right\|}, \quad \|\varphi\| \equiv \int_0^T \int_\Omega |\varphi|\, d\Omega\, dt,$$

where ψ is the streamfunction produced by integrating the model from initial conditions specified by the subscript and $\overline{\psi}^q$ denotes the solution of the tangent model linearized in the vicinity of ψ_q. As shown in Fig. 2, the correct asymptotic behavior of the Taylor expansion $\Phi(\varepsilon) \sim \varepsilon^2$, $(\varepsilon \to 0)$ is observed only with $\nu = 500\ \text{m}^2\ \text{s}^{-1}$ (i.e., when the dissipation time scale is comparable with τ_e).

Exponential growth of the small-scale harmonics in the tangent linear and adjoint codes could be suppressed by increasing the viscosity to a "stable" value ($\nu = 500\ \text{m}^2\ \text{s}^{-1}$; B. Cornuelle 2006, personal communication). This approach usually improves the performance of the adjoint 4DVAR schemes in the cases when TLA breaks down because of nonlinear instabilities. However, it does not improve the accuracy of TLA and may degrade it even further (Fig. 2). As a rule, TLA breakdown causes certain difficulties in the performance of descent algorithms, making the adjoint 4DVAR inefficient after several iterations.

c. Simulated observations

Observations ψ^*_{kn} were picked from the first-guess solutions at $t_{n=1,2,3} = 15$, 30, and 45 days at the spatial locations specified by "sparse" and "dense" measurement arrays (Fig. 1) and contaminated by white noise ε_ψ whose rms variation $\varepsilon\|\psi^{\text{ref}}\|/(\Omega T)$ was varied ($\varepsilon = 0, 0.1, 0.3$). To regularize the problem we also specified

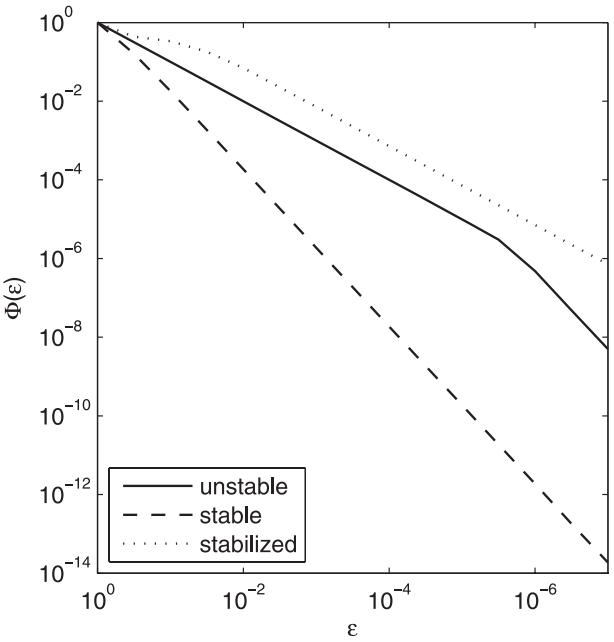

FIG. 2. TLA errors $\Phi(\varepsilon)$ of the model solutions.

the "bogus data set" $\Delta^2\psi = 0$ at $t_n = 0$, 15, 30, and 45 days. The corresponding cost function is

$$J = \frac{1}{2}\int_\Omega \left\{ \sum_{n=1}^{3}\sum_{k=1}^{K} [\hat{O}_k \psi(t_n) - \psi^*_{k,n}]^2 + W_s \sum_{n=0}^{3} [\Delta^2 \psi(t_n)]^2 \right\} d\Omega,$$

where $\hat{O}_k$ projects $\psi(t_n)$ on the kth observation point, $K = 16(64)$ is the number of observation points at time layer n, and $W_s = 0.03\delta x^4$ is the smoothing weight. The dimension of the observational space (including both real and bogus datapoints) is $n_o = 31^2 \times 4 + 3N = 3N + 3844$.

Following the notation of section 2b, the observational operator for $n = 1, 2, 3$ is represented in the matrix form:

$$\mathcal{G} = \begin{bmatrix} \sum_{k=1}^{K} \hat{O}_k \\ \sqrt{W_s}\Delta^2 \end{bmatrix} [\Delta - R_d^{-2}\mathbf{E}]^{-1},$$

where $\mathbf{E}$ is the identity matrix. For $n = 0$ the observational operator is $\mathbf{Q} = \sqrt{W_s}\Delta^2(\Delta - R_d^{-2}\mathbf{E})^{-1}$. The corresponding term of the cost function can be interpreted as the background term with $q_b = 0$ and the inverse background error covariance $\mathbf{B} = \mathbf{Q}^{\text{T}}\mathbf{Q}$.

We performed two sets of the twin-data assimilation experiments: with the stable ($\nu = 500\ \text{m}^2\ \text{s}^{-1}$) and unstable ($\nu = 50\ \text{m}^2\ \text{s}^{-1}$) adjoint models. Within each set we varied the number of observations $N = \{16, 64\}$ and

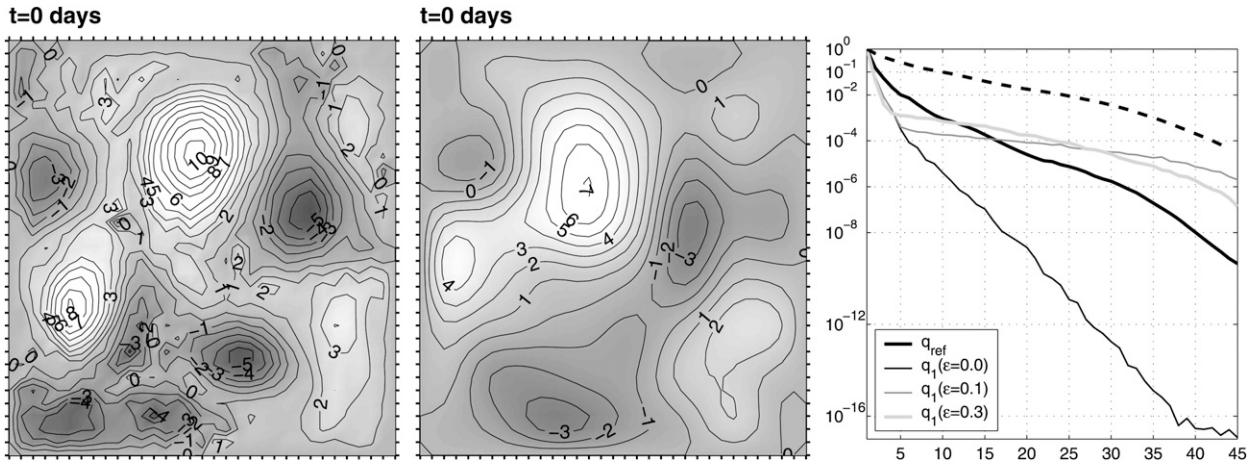

FIG. 3. Potential vorticity q of the (left) reference and (right) first-guess ($N = 64$, $\varepsilon_\psi = 0.3$) solutions at $t = 0$. The contour interval (CI) is 10^{-5} s^{-1}. (right) The normalized EOF spectra of the reference solution and of the first-guess samples retrieved from the data with different noise levels. Dashed line show the approximation error of the reference solution as a function of the number of eigenmodes used.

the noise level in the data $\varepsilon = \{0, 0.1, 0.3\}$. In the R4DVAR assimilation experiments we also varied the dimension of $\mathcal{K}^m$ $m = \{8, 15\}$ and the type of EOF decomposition in the process of generation of the Krylov spaces (section 2b). In the R4DVAR analysis the value of m is limited by the number of the time samples N. To bypass this limitation the model fields sampled at t^n were augmented with additional samples taken daily between observations.

To assess the R4DVAR performance we checked its convergence rate against 4DVAR for every set of parameters. In the unstable case ($\nu = 50$ m^2 s^{-1}) the adjoint model was stabilized by setting $\nu = 500$ m^2 s^{-1}, otherwise it was impossible to find a minimum of J with a reasonable accuracy.

The quality of reconstruction of the reference solutions was gauged by the error parameter:

$$e_\psi^2 = \frac{||\psi - \psi^{\mathrm{ref}}||}{||\psi^{\mathrm{ref}}||}.$$

d. First-guess solutions

To obtain the first-guess set of basis functions in the R4DVAR case we used the following procedure: first, the values of $\psi^*_{k,n}$, $n = 1, \ldots, 3$ were linearly interpolated on the model grid; second, the interpolated fields $\psi_k(x, y)$ were smoothed by a biharmonic filter, whose transfer function was tuned to suppress the interpolation noise and noise in the data; third, potential vorticity distributions $q_k(x, y)$ were computed as $q_k = (\Delta - R_d^{-2})\psi_k$; forth, $q_1(x, y)$ and $q_2(x, y)$ were integrated for 30 and 15 days, respectively, and subsampled with 3-day discretization; finally, $q_3(x, y)$ was added to 17 samples obtained from the integrations and the whole set of 18 fields was subjected to EOF analysis.

Figure 3 shows comparison of the q^{ref}, q_1 and the spectra EOF(q^{ref}), EOF($q_{1\ldots18}$) of the first-guess samples retrieved from the "dense" data with $\varepsilon = 0, 0.1$, and 0.3 in the unstable case. It is seen that the reference solution can be described with a relatively high precision using only 15–20 eigenmodes of the covariance function $\overline{q^{\mathrm{ref}}(x, y, t)q^{\mathrm{ref}}(x', y', t)}$, whereas the first-guess spectra differ considerably in their properties from the reference one (right panel in Fig. 3). The difference is caused by the dominance of the large-scale modes and exhibits itself in much steeper decay of the spectra. As a consequence, the first-guess solutions are well approximated by only five–seven eigenmodes of the respective covariance functions. The corresponding values of e_{ψ_1}, however, are rather large and vary in between 0.42 for $\varepsilon = 0$ and 0.51 for the $\varepsilon = 0.3$.

4. Results

We conducted a series of 60 twin-data assimilation experiments using the adjoint and R4DVAR assimilation techniques. The major purpose was to compare the convergence rates of both methods and estimate their potential ability to retrieve the reference state from the data. Results of the experiments are assembled in Tables 1 and 2.

In the R4DVAR analyses we also compared the performance of the two methods of generating the Krylov subspaces outlined in section 2b. The first one is based on EOF decomposition of the sample covariance matrix $\mathbf{C}$ (experiments labeled $E^{8,15}$), whereas the second one also accounts for the background covariance and the

TABLE 1. Results of the assimilation experiments with the stable adjoint model ($\nu = 500$ m^2 s^{-1}).

		Adjoint		E_H^8		E^8		E_H^{15}		E^{15}	
ε_ψ	δx	e_ψ	J/J_0 ($\times 10^3$)	e_ψ	J/J_0 ($\times 10^3$)	e_ψ	J/J_0 ($\times 10^3$)	e_ψ	J/J_0 ($\times 10^3$)	e_ψ	J/J_0 ($\times 10^3$)
0.0	4	0.034	0.44	0.040	1.25	0.037	1.32	0.040	1.07	0.038	1.08
	8	0.196	3.61	0.156	3.74	0.193	4.97	0.122	2.62	0.116	2.75
0.1	4	0.086	8.52	0.103	8.74	0.107	10.1	0.059	7.13	0.058	7.21
	8	0.192	11.8	0.143	5.89	0.174	7.63	0.139	5.65	0.144	5.79
0.3	4	0.181	60.1	0.145	56.2	0.164	57.8	0.136	53.8	0.136	53.9
	8	0.257	52.1	0.271	38.9	0.280	39.9	0.297	33.4	0.308	35.6

structure of the observation operators (experiments $E_H^{8,15}$). Because 4DVAR method failed to converge in the unstable case because of the breakdown of the TLN approximation, we prescribed $\nu = 500$ m^2 s^{-1} in the adjoint model for both stable and unstable runs. In 4DVAR experiments the limited-memory quasi-Newtonian descent algorithm of Byrd et al. (1995) was used. Iterations were terminated when either the relative reduction of the cost function was less than machine precision 10^{-10}, or the number of iterations exceeded 3000. In the R4DVAR experiments we performed several $\mathbf{H}^{-1}$ preconditioned iterations [Eq. (7)]. As a rule, convergence was fast, and never required more than three iterations of the internal minimization loop.

After some tuning we found that the best overall convergence rate was achieved when the control subspace was updated as soon as the gradient in the inner loop reduced more than 50 times in magnitude. This criterion was used in to compute the values listed in Tables 1 and 2, which compare e_ψ and the relative reduction of the cost function J/J_0 between the assimilation experiments. As seen from the tables, R4DVAR outperforms 4DVAR, in most cases providing better fit to the reference state and greater reduction of the cost function. The only exception is the case of dense observational array with perfect observations (first line in Table 1).

Comparing columns 2–4 and 3–5 in both tables also shows that performance of R4DVAR is better for $m = 15$ eigenfunctions. Experiments with larger m (not shown) did not improve the rate of convergence. We attribute this to the spectral properties of the reference solution, which show that q^{ref} can be approximated by 15 eigenmodes with an error of 7% (dashed line in Fig. 3c).

Differences in the values of e_ψ and J/J_0 in columns 2–3 and 4–5 indicate that E_H experiments provide a somewhat better convergence rate: the final values of J/J_0 are lower, when compared with those obtained using EOF analysis of $\mathbf{C}$. The advantage is particularly evident for $\varepsilon = 0$ and 0.1 with $m = 15$ in the unstable case and $m = 8$ in the stable case. In terms of e_ψ the difference in performance between the methods is less evident, especially for sparse observations and $\varepsilon = 0.3$. This can be partly explained by a tendency to data overfitting by the E_H method, which tends to converge faster, whereas W_s was not fine tuned to adequately account for the noise level.

Figures 4 and 5 compare convergence rates of the R4DVAR technique and the 4DVAR method. To simplify the comparison, the number of 4DVAR iterations is shown below the horizontal axis whereas the equivalent (in CPU terms) number of R4DVAR inner loop iterations is shown above the axis. Since the value of m is low, the CPU time required by EOF analysis and covariance estimation contributes only a small fraction to the total computational cost of the R4DVAR, which is almost entirely determined by the number of model runs ($m + 1$) required for gradient estimation. The adjoint code in our case required 10% more CPU time than the direct run, so that one R4DVAR iteration was approximately equivalent to 5 4DVAR iterations for $m = 8$ and 8 iterations for $m = 15$.

TABLE 2. As in Table 1, but for the unstable adjoint model ($\nu = 50$ m^2 s^{-1}).

		Adjoint		E_H^8		E^8		E_H^{15}		E^{15}	
ε_ψ	δx	e_ψ	J/J_0 ($\times 10^3$)	e_ψ	J/J_0 ($\times 10^3$)	e_ψ	J/J_0 ($\times 10^3$)	e_ψ	J/J_0 ($\times 10^3$)	e_ψ	J/J_0 ($\times 10^3$)
0.0	4	0.182	41.1	0.099	11.9	0.106	12.1	0.089	10.1	0.095	10.5
	8	0.391	56.3	0.341	29.6	0.339	29.9	0.224	14.6	0.278	20.3
0.1	4	0.242	59.5	0.128	18.5	0.135	19.3	0.129	16.8	0.125	16.9
	8	0.389	70.0	0.288	26.3	0.272	26.9	0.265	17.4	0.304	24.1
0.3	4	0.252	105.	0.174	55.1	0.182	60.0	0.167	51.2	0.168	51.4
	8	0.376	102.	0.424	62.1	0.417	64.7	0.356	46.0	0.328	47.3

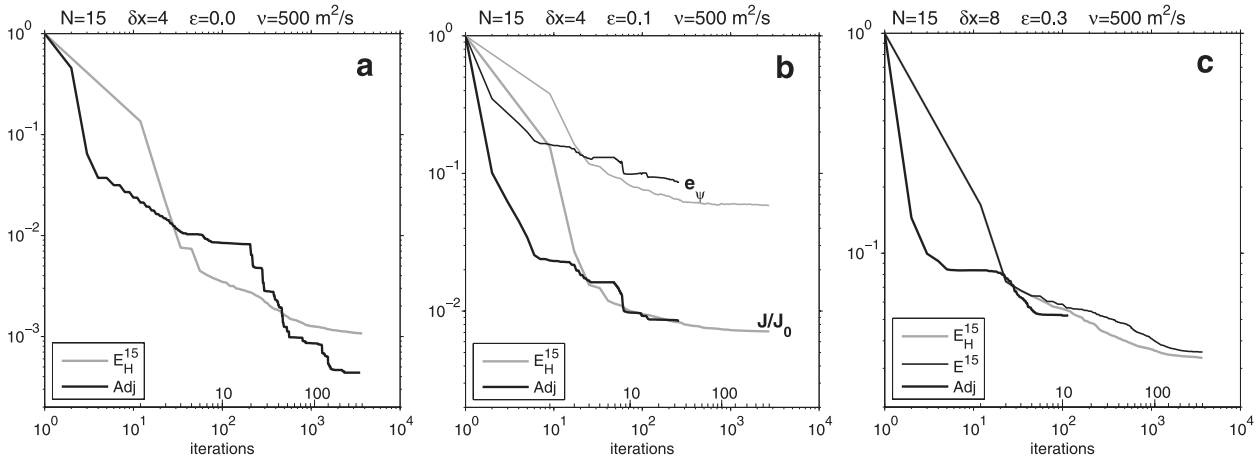

FIG. 4. Relative reduction of the cost function J/J_0 for experiments with the stable model. (middle) Evolution of e_ψ with iterations.

Figures 4 and 5 show that 4DVAR demonstrates faster convergence at the initial stages of assimilation when the number of 4DVAR iterations is $i \leq m$. However, the R4DVAR catches up after m–$4m$ iterations, performs similarly for the stable case (Fig. 4), and outperforms 4DVAR in the case when TLA is broken (Fig. 5). The effect becomes more visible at higher noise levels and sparser sampling: in these cases J has a larger number of local minima, and the 4DVAR algorithm in the stable case tends to terminate as soon as it encounters the first one (Figs. 4b,c). R4DVAR has a capability to search over the surroundings and eventually find a deeper minimum. The sparsely sampled unstable experiment with zero noise (Fig. 5c) provides an interesting example of this property: the 4DVAR scheme failed at the 76th iteration because of the loss of the descent direction, providing a "suboptimal" initial condition shown in the left panel of Fig. 6. The R4DVAR scheme proceeded further, and was able to retrieve an anticyclonic eddy in the northwestern corner of the domain (middle panel in Fig. 6). Note that this eddy is barely captured by the observational grid only at the last day of the model run (Fig. 1, rightmost panel in the top row).

As it is seen from Table 2 and Fig. 5, the R4DVAR technique is especially advantageous in the unstable case, mostly because of its robustness with respect to dynamical instabilities. In contrast, the 4DVAR algorithm terminated because of the line search failure in all the unstable cases. At sparser sampling and higher noise levels the termination occurred much earlier, often after less than 100 iterations (Figs. 5b,c).

Finally, building the Krylov subspaces using model–data projection $\mathcal{G}$ and $\mathbf{B}$ (experiments E_H) proves to be advantageous in terms of convergence rates, especially at the late stages of the assimilation process when sparse and/or noisy data are assimilated (Figs. 4c and 5). At these stages prior statistics imposed by the smoothness constraint begins to play its role, as the contribution of

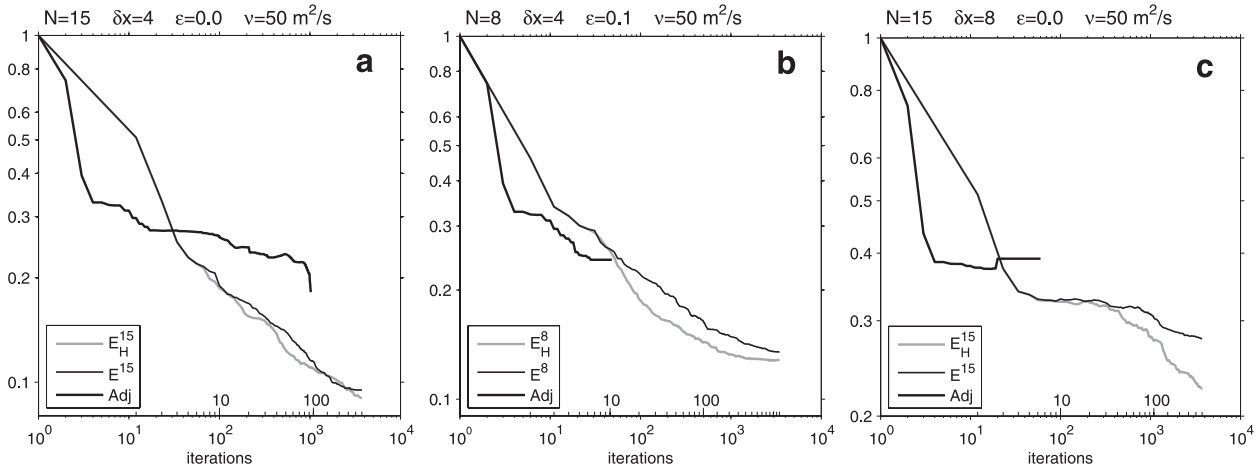

FIG. 5. Error in the approximation of the reference solution ε_ψ for experiments with the unstable model.

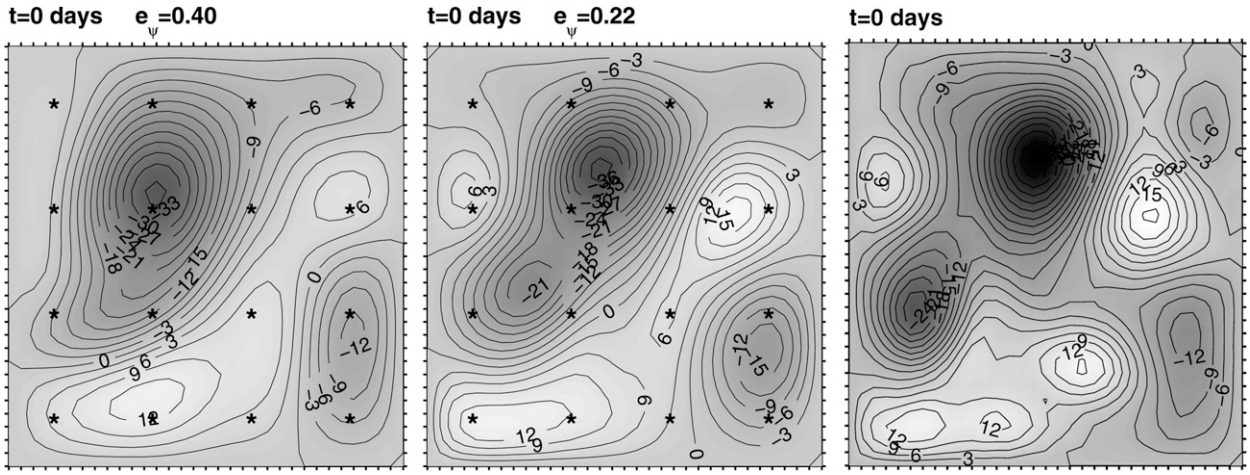

FIG. 6. Optimized streamfunction at $t = 0$ obtained by the (left) 4DVAR and (middle) R4DVAR methods for the unstable case at sparse resolution. (right) The reference solution is shown. The CI is 3000 m^2 s^{-1}.

the smoothness term to the cost function becomes comparable with the term, penalizing misfit with the real data. As a consequence, directions toward the minimum of the cost function obtained by the standard EOF expansion become more and more distorted by the ''observed'' zero values of $\Delta^2\psi$. One may expect that in the case of nonlocal model–data projection operators the effect could be visible at the earlier stages as well.

5. Discussion and conclusions

In this paper, a version of the reduced control space 4DVAR data assimilation method is proposed. In contrast to previous studies (e.g., Robert et al. 2006; Daescu and Navon 2007; Qiu et al. 2007; Liu et al. 2008), which utilized a fixed EOF-generated subspace for optimization, our algorithm employs a sequence of low-dimensional subspaces that are iteratively updated in the process of finding a minimum of the cost function. A similar optimization technique was utilized by Vermeulen and Heemink (2006), Cao et al. (2007), and Fang et al. (2009), but their approach involved construction of the reduced model and its adjoint, which are not required in our case.

The algorithm is tested in the framework of twin-data assimilation experiments with a nonlinear quasigeostrophic model controlled by the initial distribution of potential vorticity. Robustness of the method is compared with the standard 4DVAR technique based on the adjoint code. Our results can be summarized as follows:

1) Compared to 4DVAR, the proposed method provides similar or better reduction of the cost function after several updates of the search subspace. In terms of the computational cost, R4DVAR performs similarly with 4DVAR if the latter is terminated after more than $2m$–$4m$ iterations, where m is the (fixed) dimension of the reduced control subspaces (Figs. 4 and 5).
2) The proposed method gains substantial advantage over 4DVAR when the dynamical constraints have strong nonlinear instabilities, which cause the breakdown of TLA.
3) Compared to 4DVAR, the proposed method gains extra efficiency when observations become more sparse and/or noisy.

The proposed technique is based on application of the Krylov subspace method targeted on the specific type of cost functions encountered in 4DVAR problems. The technique has a lot in common with the GMRES$_m$ algorithm and appears to be equivalent to GMRES in the limit $m = M$ under the conditions specified in section 2. The distinct feature of our approach is construction of the low-dimensional subspaces not on the powers of $\mathbf{H}$, but on the approximations to the operators entering its square root in (6). Because these operators are proportional to the powers of the dynamical operator, we employ the EOF analysis of the model trajectories built on the control space residuals to extract the functions spanning the low-dimensional search spaces $\mathcal{K}^m$.

Similar to other R4DVAR methods, the proposed technique does not require development and maintenance of the adjoint code. It is also very efficient in terms of parallelization, since the major portion of CPU time is consumed by m-independent model runs required for gradient computation in the Krylov subspaces. Regarding parallelization, one may expect an additional 10%–30% gain in computational cost of the R4DVAR method when it is applied to state-of-the-art OGCMs, whose parallelization efficiency scales nonlinearly with an increase of the number of processors.

Another potential CPU time gain in OGCM applications should be acquired if we consider computational efficiency of the adjoint models. As a rule, adjoint codes of the community OGCMs with 4DVAR capabilities (especially those generated by automatic compilers) require 2–5 times more CPU time than the direct codes, whereas the adjoint of our simple model is only 10% slower than the direct one.

One particular advantage of the R4DVAR approach is that it eliminates the necessity to ''stabilize'' the adjoint model in the presence of nonlinear instabilities (e.g., Zhu and Kamachi 2000; Zhu et al. 2002) by simplification and/or modification of the numerical scheme. These approximations lead to a certain loss of accuracy of the tangent linear approximation (Fig. 3) and degrade the performance of descent algorithms.

Another benefit of the method is that it implicitly regularizes the problem through the ranking of the search subspaces in the course of assimilation: during the first iterations the smoothest approximations to initial conditions are recovered, they are later refined in subspaces containing higher-order spatial harmonics. In fact, results of R4DVAR assimilation were weakly sensitive to the magnitude of W_s, especially in the cases of dense observations and moderate noise levels.

Our numerical experiments have also shown that the efficiency of the proposed technique is sensitive to the quality of the first-guess subspace $\mathcal{K}_0$. For instance, when $\mathcal{K}_0$ is generated by an ensemble of white noise perturbations as proposed by Qiu et al. (2007), the descent became inefficient and required an order in magnitude increase either in the subspace dimension or in the number of iterations needed to achieve the convergence. This phenomenon can be explained by a small projection of the seed perturbations on the optimal state. A similar effect has been observed in the twin-data experiments of Qiu et al. 2007, who considered a smooth background state of a 2D shallow-water model on the 44×44 grid but had to use 150 ensemble members for a reasonably accurate assimilation.

In that respect it is necessary to note that in the limiting case of ''zero quality'' of the first-guess state ($\mathbf{x}_0$ is $\mathbf{H}$-orthogonal to $\mathbf{b}$, section 2d), the proposed algorithm will fail in the linear case. This is not necessarily true in the nonlinear case, because in the process of model integration nonlinearities may generate state vector components that are not present in the first-guess solution. These components affect the composition of $\mathcal{K}^m$ and may eventually span $\mathbf{b}$ with iterations. In operational applications the very bad quality of the background state seems to be unlikely, because a reasonably good approximation to reality is already available either from previous assimilation cycles or from the preliminary data analysis (simulated in section 3d). Note that in all the reported R4DVAR experiments we did not use any prior information on the solution except that contained in the data itself and in a simple smoothness constraint.

Further improvements of the method can be done in several directions. First of all, a better approximation to $\mathbf{S}^{\mathrm{T}}$ could be developed. In this study we used the background error covariance to project $\mathcal{G}_k^{\mathrm{T}}\mathcal{G}_k\mathbf{A}^k\mathbf{r}$ on the state space. In many applications, however, the background error covariance is rank deficient as it is approximated by $n \sim 50$–100 leading eigenmodes, emerging from statistical analysis. In such situations the condition of Hayami et al. (2007) may be violated, causing inability of the scheme to retrieve data components not present in the spectrum of $\mathbf{B}$. More secure adjointless projections could be suggested, that involve replacing $\mathbf{A}^{\mathrm{T}}$ by $\mathbf{A}$ or by its ''nonlinear approximation.'' The latter can be obtained, for example, by reversing the sign of the odd-order differential operators in $\mathbf{A}$. These projections may seem more robust as compared to the one with low-rank B because dimensions of their null spaces are much smaller than $M - n$, and their structure may differ only marginally from the null space of $\mathbf{A}^{\mathrm{T}}$. Another possible improvement that could be done is augmenting the Krylov matrices with projections of the residuals on the certain eigenfunctions of $\mathbf{B}$, if the latter are readily available.

There is also a room for increasing the computational efficiency of the proposed algorithm. One of the directions is applying more sophisticated methods for extracting the basis in $\mathcal{K}^m$, such as direct SVD decomposition of the Krylov matrices. Another improvement could be obtained by adaptive adjustment of the Krylov space dimensions. One of the possible strategies in that respect is the entropy analysis of the Hessian spectra in $\mathcal{K}^m$ (Uzunoglu et al. 2007).

The most urgent development, however, is to test the method with the multivariate state vectors of a state-of-the-art OGCM. This is the subject of our present research.

Acknowledgments. This study was supported by the Japan Agency for Marine-Earth Science and Technology (JAMSTEC), by NASA through Grant NNX07AG53G, and by NOAA through Grant NA17RJ1230 through their sponsorship of research activities at the International Pacific Research Center. Nechaev and Panteleev were supported by the NSF Award ARC-0629400. The authors would like to thank the anonymous reviewers for their comments and suggestions, which helped to improve this paper. This work was partly sponsored by the Office of Naval Research Program Element 0601153N as part of the project ''Variational Data Assimilation for Ocean Prediction.''

REFERENCES

Arnoldi, W. E., 1951: The principle of minimized iterations in the solution of the matrix eigenvalue problem. *Quart. Appl. Math.,* **9,** 17–29.

Bennett, A. F., 1992: *Inverse Methods in Physical Oceanography*. Cambridge University Press, 346 pp.

Blessing, S., R. Greatbatch, K. Fraedrich, and F. Lunkeit, 2008: Interpreting the atmospheric circulation trend during the last half of the twentieth century: Application of an adjoint model. *J. Climate,* **21,** 4629–4646.

Byrd, R. H., P. Lu, and J. Nocedal, 1995: A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.,* **16,** 1190–1208.

Cao, Y., J. Zhu, I. M. Navon, and Z. Luo, 2007: A reduced-order approach to four-dimensional variational data assimilation using proper orthogonal decomposition. *Int. J. Numer. Methods Fluids,* **53,** 1571–1583.

Daescu, D. N., and I. M. Navon, 2007: Efficiency of a POD-based reduced second order adjoint model in 4D-Var data assimilation. *Int. J. Numer. Methods Fluids,* **53,** 985–1004.

Di Lorenzo, E., A. M. Moore, H. G. Arango, B. D. Cornuelle, A. J. Millerd, B. Powell, B. S. Chua, and A. F. Bennett, 2007: Weak and strong constraint data assimilation in the inverse Regional Ocean Modeling System (ROMS): Development and application for a baroclinic coastal upwelling system. *Ocean Modell.,* **16,** 160–187.

Evensen, G., 2003: The ensemble Kalman filter: Theoretical formulation and practical implementation. *Ocean Dyn.,* **53,** 343–357, doi:10.1007/s10236-003-0036-9.

——, 2006: *Data Assimilation: The Ensemble Kalman Filter*. Springer, 288 pp.

Fang, F., C. C. Pain, I. M. Navon, G. J. Gorman, M. D. Piggott, P. A. Allison, P. E. Farell, and A. H. Goddard, 2009: A POD-reduced order unstructured mesh ocean modelling method for moderate Reynolds number flows. *Ocean Modell.,* **28,** 127–136.

Farrell, B. F., and P. J. Ioannou, 2001: Accurate low-dimensional approximation of linear dynamics of fluid flow. *J. Atmos. Sci.,* **58,** 2771–2789.

——, and ——, 2006: Approximating optimal state estimation. *Predictability of Weather and Climate,* T. N. Palmer and R. Hagedorn, Eds., Cambridge University Press, 181–216.

Hayami, K., J.-F. Yin, and T. Ito, 2007: GMRES methods for least squares problems. NII Tech. Rep. 2007–009E, Tokyo, Japan, 28 pp.

Le Dimet, F.-X., and O. Talagrand, 1986: Variational algorithms for analysis and assimilation of meteorological observations: Theoretical aspects. *Tellus,* **38,** 97–110.

Liu, C., Q. Xiao, and B. Wang, 2008: An ensemble-based four-dimensional variational data assimilation scheme. Part I: Technical formulation and preliminary test. *Mon. Wea. Rev.,* **136,** 3363–3373.

Oldenborgh, G. J., G. Burgers, S. Ventzke, C. Eckert, and R. Giering, 1999: Tracking down the ENSO delayed oscillator with an adjoint OGCM. *Mon. Wea. Rev.,* **127,** 1477–1496.

Ott, E., B. R. Hunt, I. Szuyogh, A. V. Zimin, E. J. Kostelich, M. Corazza, E. Kalnay, and D. J. Patil, 2004: A local ensemble Kalman filter for atmospheric data assimilation. *Tellus,* **56A,** 415–428.

Pannekoucke, O., and S. Massart, 2008: Estimation of the local diffusion tensor and normalization for heterogeneous correlation modelling using a diffusion equation. *Quart. J. Roy. Meteor. Soc.,* **134,** 1425–1438.

Qiu, C., A. Shao, and L. Wei, 2007: Fitting model fields to observations by using singular value decomposition: An ensemble-based 4DVar approach. *J. Geophys. Res.,* **112,** D11105, doi:10.1029/2006JD007994.

Robert, C., S. Durbiano, E. Blayo, J. Verron, J. Blum, and F.-X. Le Dimet, 2005: A reduced-order strategy for 4D-Var data assimilation. *J. Mar. Syst.,* **57** (1–2), 70–82.

——, E. Blayo, and J. Verron, 2006: Reduced-order 4D-Var: A preconditioner for the incremental 4D-Var data assimilation method. *Geophys. Res. Lett.,* **33,** L18609, doi:10.1029/2006GL026555.

Saad, Y., 2003: *Iterative Methods for Sparse Linear Systems*. 2nd ed. SIAM, 528 pp.

Stammer, D., and Coauthors, 2003: Volume, heat and freshwater transports of the global ocean circulation 1993–2000, estimated from a general circulation model constrained by World Ocean Circulation Experiment (WOCE) data. *J. Geophys. Res.,* **108,** 3007, doi:10.1029/2001JC001115.

Talagrand, O., and F. Bouttier, 2009: *Data Assimilation in Meteorology and Oceanography*. Academic Press, in press.

Thacker, W. C., 1988: Fitting models to inadequate data by enforcing spatial and temporal smoothness. *J. Geophys. Res.,* **93,** 10 556–10 566.

Uzunoglu, B., C. J. Fletcher, M. Zupanski, and I. M. Navon, 2007: Adaptive ensemble reduction and inflation. *Quart. J. Roy. Meteor. Soc.,* **133,** 1281–1294.

Vermeulen, P. T. M., and A. W. Heemink, 2006: Model-reduced variational data assimilation. *Mon. Wea. Rev.,* **134,** 2888–2899.

Weaver, A., and P. Courtier, 2001: Correlation modeling on a sphere using a generalized diffusion equation. *Quart. J. Roy. Meteor. Soc.,* **127,** 1815–1846.

Wenzel, M., J. Schröter, and D. Olbers, 2001: The annual cycle of the global ocean circulation as determined by 4DVar data assimilation. *Prog. Oceanogr.,* **48,** 73–119.

Wunsch, C., 1996: *The Ocean Circulation Inverse Problem*. Cambridge University Press, 442 pp.

Yaremchuk, M., 2006: Sea surface salinity constrains rainfall estimates over tropical oceans. *Geophys. Res. Lett.,* **33,** L15605, doi:10.1029/2006GL026582.

Zhu, J., and M. Kamachi, 2000: The role of time step size in numerical stability of tangent linear models. *Mon. Wea. Rev.,* **128,** 1562–1572.

——, W. Hui, and M. Kamachi, 2002: The improvement made by a modified TLM in 4DVar with a geophysical boundary layer model. *Adv. Atmos. Sci.,* **19** (4), 563–582.

Zupanski, M., 2005: Maximum likelihood ensemble filter: Theoretical aspects. *Mon. Wea. Rev.,* **133,** 1710–1726.

——, D. Zupanski, T. Vukicevic, K. Eis, and T. I. V. Haar, 2005: CIRA/CSU four-dimensional variational data assimilation system. *Mon. Wea. Rev.,* **133,** 829–843.